Journal of Machine Learning Research 26 (2025) 1-29 Submitted 4/23; Revised 8/24; Published 5/25

Bayesian Data Sketching for Varying Coefficient Regression Models

Rajarshi Guhaniyogi RAJGUHANIYOGI@TAMU.EDU
Department of Statistics
Texas A & M University
College Station, TX 77843-3143, USA

Laura Baracaldo LNBARACALDOL@UCSB.EDU
Department of Statistics and Applied Probability
University of California Santa Barbara
Santa Barbara, CA 93106-3110, USA

Sudipto Banerjee SUDIPTO@UCLA.EDU
UCLA Department of Biostatistics
University of California Los Angeles
Los Angeles, CA 90095-1772, USA.

Editor: Ryan Adams

Abstract

Varying coefficient models are popular for estimating nonlinear regression functions in functional data models. Their Bayesian variants have received limited attention in large data applications, primarily due to prohibitively slow posterior computations using Markov chain Monte Carlo (MCMC) algorithms. We introduce Bayesian data sketching for varying coefficient models to obviate computational challenges presented by large sample sizes. To address the challenges of analyzing large data, we compress the functional response vector and predictor matrix by a random linear transformation to achieve dimension reduction and conduct inference on the compressed data. Our approach distinguishes itself from several existing methods for analyzing large functional data in that it requires neither the development of new models or algorithms nor any specialized computational hardware while delivering fully model-based Bayesian inference. Well-established methods and algorithms for varying-coefficient regression models can be applied to the compressed data. We establish posterior contraction rates for estimating the varying coefficients and predicting the outcome at new locations with the randomly compressed data model. We use simulation experiments and analyze remote sensed vegetation data to empirically illustrate the inferential and computational efficiency of our approach.

Keywords: B-splines, Predictive Process, Posterior contraction, Random compression matrix, Varying coefficient models.

1. Introduction

We develop a statistical learning framework for functional data analysis using Bayesian data sketching to deliver inference that scales massive functional datasets. "Data sketching" (Vempala, 2005; Halko et al., 2011; Mahoney, 2011; Woodruff, 2014; Guhaniyogi and Dunson, 2015, 2016) is a compression method that is increasingly used to analyze massive

amounts of data. The entire data set is compressed before being analyzed for computational efficiency. Data sketching proceeds by transforming the original data through a random linear transformation to produce a much smaller number of data samples. We analyze the compressed data, thereby achieving dimension reduction.

Such developments have focused mainly on ordinary linear regression and penalized linear regression (Zhang et al., 2013; Chen et al., 2015; Dobriban and Liu, 2018; Drineas et al., 2011; Ahfock et al., 2017; Huang, 2018), we develop such methods for functional regression models. Our primary challenge is probabilistic learning for the underlying effects of functional coefficients in the context of varying regression models. Although we have some similarities, our current contribution differs from compressed sensing (Donoho, 2006; Ji et al., 2008; Candes and Tao, 2006; Eldar and Kutyniok, 2012; Yuan et al., 2014) in inferential objectives. Specifically, compressed sensing solves an inverse problem by "nearly" recovering a sparse vector of responses from a smaller set of random linear transformations. In contrast, our functionally indexed response vector is not necessarily sparse. In addition, we do not seek to recover (approximately) the original values in the response vector.

We consider a varying-coefficient model (VCM) where all functional variables (response and predictors) are defined in a d-dimensional indexed space $\mathcal{D} \subseteq \mathbb{R}^d$. For temporal data $d = 1$ and for spatial data applications $d = 2$, while for spatial-temporal applications the domain is $\mathcal{D} = \mathbb{R}^2 \times \mathbb{R}^+$ and the index is a space-time tuple ($\boldsymbol{u} = (\boldsymbol{s}, t)$). For each index $\boldsymbol{u} \in \mathcal{D}$, the functional response $y(\boldsymbol{u}) \in \mathcal{Y} \subseteq \mathbb{R}$ and P functional predictors $x_1(\boldsymbol{u}), ..., x_P(\boldsymbol{u}) \in \mathcal{X} \subseteq \mathbb{R}$, are related according to a posited varying coefficients regression model

$$y(\boldsymbol{u}) = \sum_{j=1}^{P} x_j(\boldsymbol{u})\beta_j + \sum_{j=1}^{\tilde{P}} \tilde{x}_j(\boldsymbol{u}) w_j(\boldsymbol{u}) + \epsilon(\boldsymbol{u}) = \boldsymbol{x}(\boldsymbol{u})^{\mathrm{T}}\boldsymbol{\beta} + \tilde{\boldsymbol{x}}(\boldsymbol{u})^{\mathrm{T}}\boldsymbol{w}(\boldsymbol{u}) + \epsilon(\boldsymbol{u}) \,, \qquad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^{\mathrm{T}}$ is a $P \times 1$ vector of functionally static coefficients, $\tilde{\boldsymbol{x}}(\boldsymbol{u}) = (\tilde{x}_1(\boldsymbol{u}), \tilde{x}_2(\boldsymbol{u}), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}))^{\mathrm{T}}$ is a $\tilde{P} \times 1$ vector comprising a subset of predictors from $\boldsymbol{x}(\boldsymbol{u})$ (so $\tilde{P} \leq P$) whose impact on the response is expected to vary over the functional inputs, $\boldsymbol{w}(\boldsymbol{u}) = (w_1(\boldsymbol{u}), w_2(\boldsymbol{u}), \ldots, w_{\tilde{P}}(\boldsymbol{u}))^{\mathrm{T}}$ is a $\tilde{P} \times 1$ vector of functionally varying regression slopes, and $\epsilon(\boldsymbol{u}) \overset{iid}{\sim} N(0, \sigma^2)$ captures measurement error variation at location $\boldsymbol{u}$. Functionally varying regression coefficient models are effective in learning the varying impact of predictors on the response in time series (see, e.g., Chen and Tsay, 1993; Cai et al., 2000, and references therein) and in spatial applications (see, e.g., Gelfand et al., 2003; Banerjee and Johnson, 2006; Wheeler and Calder, 2007; Finley et al., 2011; Guhaniyogi et al., 2013; Finley and Banerjee, 2020; Kim and Wang, 2021, and references therein) and in spatial-temporal data analysis (see, e.g., Lee et al., 2021, and references therein). When $d = 2$, customary geostatistical regression models with only a spatially-varying intercept emerge if the first column of $x(\boldsymbol{u})$ is the intercept and $\tilde{P} = 1$ with $\tilde{x}_1(\boldsymbol{u}) = 1$. Spatially varying coefficient models, a class of varying coefficient models for $d = 2$, also offer a process-based alternative to the widely used geographically weighted regression (see, e.g., Brunsdon et al., 1996) to model non-stationary behavior in the mean. Finley (2011) offers a comparative analysis and highlights the richness of (1) in ecological applications.

Bayesian inference for (1) is computationally expensive for large data sets due to the high-dimensional covariance matrix introduced by $w(\boldsymbol{u})$ in (1). The modeling of high-dimensional dependent functional data has been attracting significant interest, and the

growing literature on scalable methods, which has been adapted and built on scalable spatial models (see, e.g., Banerjee, 2017; Heaton et al., 2019, for reviews in spatial statistics), is too vast to be comprehensively reviewed here. Briefly, model-based dimension reduction in functional data models have proceeded from fixed-rank representations (e.g., Cressie and Johannesson, 2008; Banerjee et al., 2008; Wikle, 2010; Snelson and Ghahramani, 2005; Burt et al., 2020), multi-resolution approaches (e.g., Nychka et al., 2015; Guhaniyogi and Sansó, 2018), sparsity-inducing processes (e.g., Vecchia, 1988; Datta et al., 2016; Zhang et al., 2019; Katzfuss and Guinness, 2021; Peruzzi et al., 2022) and divide-and-conquer approaches such as meta-kriging (Guhaniyogi and Banerjee, 2018; Guhaniyogi et al., 2020a,b).

While most of the aforementioned methods entail new classes of models and approximations, or very specialized high-performance computing architectures, Bayesian data sketching has the advantage that customary exploratory data analysis tools, well-established methods, and well-tested available algorithms for implementing (1) can be applied to the sketched data without requiring new algorithmic or software development. We pursue fully model-based Bayesian data sketching, where inference proceeds from a hierarchical model (Cressie and Wikle, 2015; Banerjee et al., 2014). The hierarchical approach to functional data analysis is widely employed for inferring on model parameters that may be weakly identified from the likelihood alone and, more relevantly for substantive inference, for estimating the functional relationship between response and predictors over the domain of interest. For analytic tractability, we model the varying coefficients using basis expansions (Wikle, 2010; Wang et al., 2008; Wang and Xia, 2009; Bai et al., 2019) rather than Gaussian processes.

We exploit some recent developments in the theory of random matrices to relate the inference from the compressed data with the full-scale functional data model. We establish consistency of the posterior distributions of the varying coefficients and analyze the predictive efficiency of our models based on compressed data. Posterior contraction of varying-coefficient (VC) models have been investigated by a few recent articles. For example, Guhaniyogi et al. (2020b) derive minimax optimal posterior contraction rates for Bayesian VC models under GP priors when the number of predictors P is fixed. Deshpande et al. (2020) also derived near-optimal posterior contraction rates under BART priors, and Bai et al. (2019) provided an asymptotically optimal rate of estimation for varying coefficients with a variable selection prior on varying coefficients. We address these questions in the context of data compression, which has largely remained unexplored.

While our approach randomly compresses the data for efficient Bayesian inference, there is a related but distinct approach that relies on stochastic gradient decent with subsampled or mini-batch input at each iteration for efficient computation. Traditionally, both proposal generation and acceptance test within the Metropolis-Hastings algorithm require a full pass over the data, which results in reduced efficiency. Subsampling approaches address this by using mini-batches or subsets of data for both the proposal step and the acceptance test. For the proposal, Welling and Teh (2011) introduced the Stochastic Gradient Langevin Dynamics (SGLD) algorithm, a variant of the first-order Langevin dynamics that adds noise to ensure the correct noise distribution. They also anneal the step size to zero, eliminating the need for an acceptance test. Ahn et al. (2012) moved away from Langevin dynamics and proposed a method based on Fisher scoring. Chen et al. (2014) introduced the Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) algorithm, which is based on a variant of second-order Langevin dynamics with momentum to update the state. Similarly

to SGLD, SGHMC injects additional noise, but also reduces the effect of gradient noise. For the acceptance test, Korattikara et al. (2014) proposed a sequential hypothesis test for Metropolis-Hastings proposals based on a fraction of the full dataset. Building on this seminal work, other mini-batch MH algorithms were developed by Seita et al. (2016) and Bardenet et al. (2014). In recent years, mini-batched approaches in Markov chain Monte Carlo (MCMC) have expanded to include tempered methods (Li and Wong, 2017), Gibbs sampling (De Sa et al., 2018), and gradient-based proposals (Wu et al., 2022), with a comprehensive review provided by Bardenet et al. (2017). Subsampling-based approaches are further extended to derive distributed Bayesian approaches, where instead of computing a gradient with different subsamples at each step, posterior distributions are independently fitted on different mini-batches followed by combining the inferences from these mini-batches (Guhaniyogi and Banerjee, 2018, 2019; Guhaniyogi et al., 2020b, 2023). Here, we build the compressed model and are agnostic to the specific estimation algorithm. In fact, while we use MCMC in subsequent analysis, alternative approaches such as predictive stacking (Zhang et al., 2024) can be used to learn about the functional coefficients.

The balance of this article proceeds as follows. Section 2 develops our data sketching approach and discusses Bayesian implementation of VC models with sketched data. Section 3 establishes posterior contraction rates for varying coefficients under data sketching. Section 4 demonstrates performance of the proposed approach with simulation examples and a forestry data analysis. Finally, Section 6 concludes the paper with an eye toward future extensions. All proofs of the theoretical results are placed in the Appendix.

2. Bayesian Compressed Varying Coefficient Models

2.1 Model construction

We model each varying coefficient $w_j(\boldsymbol{u})$ in (1) as

$$w_j(\boldsymbol{u}) = \sum_{h=1}^{H} B_{jh}(\boldsymbol{u})\gamma_{jh} \ , \quad j = 1, ..., \tilde{P} \ , \tag{2}$$

where each $B_{jh}(\boldsymbol{u})$ is a basis function evaluated at an index $\boldsymbol{u}$ for $h = 1, ..., H$, and γ_{jh}'s are the corresponding basis coefficients. The distribution of these γ_{jh}'s yields a multivariate process with $\text{cov}(w_i(\boldsymbol{u}), w_j(\boldsymbol{u}')) = \boldsymbol{B}_i(\boldsymbol{u})^{\text{T}}\text{cov}(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j)\boldsymbol{B}_j(\boldsymbol{u})$, where $\boldsymbol{B}_i(\boldsymbol{u})$ and $\boldsymbol{\gamma}_i$ are $H \times 1$ with elements $B_{ih}(\boldsymbol{u})$ and γ_{ih}, respectively, for $h = 1, \ldots, H$.

Appropriate basis functions can produce appropriate classes of multivariate functional processes. Several choices are available. For example, Biller and Fahrmeir (2001) and Huang et al. (2015) use splines to model the $B_{jh}(\boldsymbol{u})$'s and place Gaussian priors on the basis coefficients γ_{jh}. Li et al. (2015) propose a scale-mixture of multivariate normal distributions to shrink groups of basis coefficients toward zero. More recently, Bai et al. (2019) proposed using B-spline basis functions and multivariate spike-and-slab discrete mixture prior distributions on basis coefficients to aid selection of functional variables. Other popular choices for basis functions include wavelet basis (Vidakovic, 2009; Cressie and Wikle, 2015), radial basis (Bliznyuk et al., 2008), and locally bisquare (Cressie and Johannesson, 2008) or elliptical basis functions (Lemos and Sansó, 2009). Alternatively, a basis representation of $w_j(\boldsymbol{u})$ can be constructed by envisioning $w_j(\boldsymbol{u})$ as the projection of a Gaussian process $w_j(\boldsymbol{u})$ onto

a set of reference points, or "knots", producing predictive processes or sparse Gaussian processes (Snelson and Ghahramani, 2005; Banerjee et al., 2008; Guhaniyogi et al., 2013, see, e.g.,). More generally, each $w_j(\boldsymbol{u})$ can be modeled using multiresolution analogues of the aforementioned models to capture global variations at lower resolution and local variations at higher resolution (Katzfuss, 2017; Guhaniyogi and Sansó, 2018).

Let $\{y(\boldsymbol{u}_i), \boldsymbol{x}(\boldsymbol{u}_i)\}$ be observations at the N index points $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N\}$. Using (2) in (1) yields the Gaussian linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{X}}\boldsymbol{B}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\ , \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_N)\ . \tag{3}$$

where $\boldsymbol{y} = (y(\boldsymbol{u}_1), y(\boldsymbol{u}_2), \ldots, y(\boldsymbol{u}_N))^{\mathrm{T}}$ and $\boldsymbol{\epsilon} = (\epsilon(\boldsymbol{u}_1), \epsilon(\boldsymbol{u}_2), \ldots, \epsilon(\boldsymbol{u}_N))^{\mathrm{T}}$ are $N \times 1$ vectors of responses and errors, respectively, $\boldsymbol{X}$ is $N \times P$ with n-th row $\boldsymbol{x}(\boldsymbol{u}_n)^{\mathrm{T}}$, $\tilde{\boldsymbol{X}}$ is the $N \times N\tilde{P}$ block-diagonal matrix with (n,n)-th block $\tilde{\boldsymbol{x}}(\boldsymbol{u}_n)^{\mathrm{T}}$, $\boldsymbol{B} = (\boldsymbol{B}(\boldsymbol{u}_1)^{\mathrm{T}}, \ldots, \boldsymbol{B}(\boldsymbol{u}_N)^{\mathrm{T}})^{\mathrm{T}}$ is $N\tilde{P} \times H\tilde{P}$ with $\boldsymbol{B}(\boldsymbol{u}_n)$ a block-diagonal $\tilde{P} \times H\tilde{P}$ matrix whose j-th diagonal block is $(B_{j1}(\boldsymbol{u}_n), \ldots, B_{jH}(\boldsymbol{u}_n))$. The coefficient $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, ..., \boldsymbol{\gamma}_{\tilde{P}}^{\mathrm{T}})^{\mathrm{T}}$ is $H\tilde{P} \times 1$ with each $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jH})^{\mathrm{T}}$ being $H \times 1$. Bayesian methods for estimating (3) typically employ a multivariate normal prior (Biller and Fahrmeir, 2001; Huang et al., 2015) or its scale-mixture (discrete as well as continuous) variants (Li et al., 2015; Bai et al., 2019) on $\boldsymbol{\gamma}$.

While the basis functions project the coefficients into a low-dimensional space, working with (3) will be still be expensive for large N and will be impracticable for delivering full inference (with robust probabilistic uncertainty quantification) for data sets with $N \sim 10^5+$ on modest computing environments. Furthermore, as is well understood in linear regression, specifying a small number of basis functions in (3) can lead to substantial over-smoothing and, consequently, biased residual variance estimates in functional varying coefficient models (see, e.g., the discussion in Section 2.1 of Banerjee, 2017, including Figures 1 and 2 in the paper). Instead, we consider data compression or sketching using a random linear mapping to reduce the size of the data from N to M observations. For this, we use M one-dimensional linear mappings of the data encoded by an $M \times N$ compression matrix $\boldsymbol{\Phi}$ with $M << N$. This compression matrix is applied to $\boldsymbol{y}$, $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ to construct the $M \times 1$ compressed response vector $\boldsymbol{y}_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\boldsymbol{y}$ and the matrices $\boldsymbol{X}_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\tilde{\boldsymbol{X}}$. We will return to the specification of $\boldsymbol{\Phi}$, which, of course, will be crucial for relating the inference from the compressed data with the full model. For now assuming that we have fixed $\boldsymbol{\Phi}$, we construct a Bayesian hierarchical model for the compressed data

$$p(\boldsymbol{\psi}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 \,|\, \boldsymbol{y}_{\boldsymbol{\Phi}}, \boldsymbol{\Phi}) \propto p(\boldsymbol{\psi}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}) \times N(\boldsymbol{y}_{\boldsymbol{\Phi}} \,|\, \boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta} + \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\gamma}, \sigma^2 I_M)\ , \tag{4}$$

where $\boldsymbol{\psi}$ denotes additional parameters specifying the prior distributions on either $\boldsymbol{\gamma}$ or $\boldsymbol{\beta}$. For example, a customary specification is

$$p(\boldsymbol{\psi}, \sigma^2, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{\tilde{P}} IG(\tau_i^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta) \times N(\boldsymbol{\gamma} \,|\, \boldsymbol{0}, \boldsymbol{\Delta})\ , \tag{5}$$

where $\boldsymbol{\psi} = \{\tau_1^2, ..., \tau_{\tilde{P}}^2\}$ and $\boldsymbol{\Delta}$ is $H\tilde{P} \times H\tilde{P}$ block-diagonal with j-th block given by $\tau_j^2 \boldsymbol{I}_H$, for $j = 1, ..., \tilde{P}$. While (5) is a convenient choice for empirical investigations due to conjugate full conditional distributions, our method applies broadly to any basis function and any discrete or continuous mixture of Gaussian priors on the basis coefficients. In applications

where the associations among the latent regression slopes is of importance, one could, for instance, adopt $p(\boldsymbol{\psi}, \boldsymbol{\gamma}) = IW(\boldsymbol{\psi} \,|\, r, \boldsymbol{\Omega}) \times N(\boldsymbol{\gamma} \,|\, 0, \boldsymbol{\Delta}_{\boldsymbol{\psi}})$ with $\boldsymbol{\psi}$ as the $H\tilde{P} \times H\tilde{P}$ covariance matrix for $\boldsymbol{\gamma}$. Our current focus is not, however, on such multivariate models, so we do not discuss them further except to note that (4) accommodates such extensions.

The likelihood in (4) is different from that by applying $\boldsymbol{\Phi}$ to (3) because the error distribution in (4) is retained as the usual noise distribution without any effect of $\boldsymbol{\Phi}$. Hence, the model in (4) is a model analogous to (3) but applied to the *new* compressed data $\{\boldsymbol{y}_{\boldsymbol{\Phi}}, \boldsymbol{X}_{\boldsymbol{\Phi}}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\}$. However, (4) can be regarded as an approximately compressed version of (3) when $\boldsymbol{\Phi}$ is a random matrix constructed in a manner customary for sketching matrices (Sarlos, 2006). To see this, note that a compressed version of Equation (3) will lead to an error $\boldsymbol{\Phi\epsilon}$ which follows $\mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{\Phi}\boldsymbol{\Phi}^T)$. Lemma 5.36 and Remark 5.40 of Vershynin (2010) ensure that when $\boldsymbol{\Phi}$ is a random matrix constructed as described in this article, the condition $||\boldsymbol{\Phi}\boldsymbol{\Phi}^T - \boldsymbol{I}|| \leq C'\sqrt{M/N}$ for some constant C' is met with probability at least $1 - \exp(-C''M)$. Hence, with a very high probability, $\boldsymbol{\Phi\epsilon}$ behaves approximately as an M-dimensional i.i.d. noise when M/N is small, thus building a connection between Model (3) and (4). This connection is also key to the computational benefits offered by our model, since working with a $\boldsymbol{\Phi}$-transformed model (3), where the noise distribution is transformed according to $\boldsymbol{\Phi\epsilon}$, will not deliver the computational benefits we desire.

For specifying $\boldsymbol{\Phi}$ we pursue "data oblivious Gaussian sketching" (Sarlos, 2006), where we draw the elements of $\boldsymbol{\Phi} = (\Phi_{ij})$ independently from $\mathrm{N}(0, 1/N)$ and fix them. The dominant computational operations for obtaining the sketched data using Gaussian sketches is $O(MN^2\tilde{P})$. While Gaussian sketching constructs dense matrices, there are alternative options, oblivious to data, such as the Hadamard sketch (Ailon and Chazelle, 2009) and the Clarkson - Woodruff sketch (Clarkson and Woodruff, 2017) that are available for $\boldsymbol{\Phi}$. These strategies employ discrete distributions (e.g., Rademacher distribution), instead of a Gaussian distribution, to construct sparse random matrices, which enhances computational efficiency for sketching large data matrices. However, this is less crucial in Bayesian settings since the computation time of (4) far exceeds that for the construction of the sketched data. The compressed data serves as a surrogate for the Bayesian regression analysis with varying coefficients. Since the number of compressed records is much smaller than the number of records in the uncompressed data matrix, model fitting becomes computationally efficient and economical in terms of storage as well as the number of floating point operations (flops). Importantly, the original data are not recoverable from the compressed data, and the latter reveal no more information than would be revealed by a completely new sample (Zhou et al., 2008). In fact, the original uncompressed data does not need to be stored or accessed at any stage in the course of the analysis.

2.2 Posterior Computations & Predictive Inference

In what follows, we discuss efficient computation offered by the data sketching framework. With prior distributions on parameters specified as in (5), posterior computation requires drawing Markov chain Monte Carlo (MCMC) samples sequentially from the full conditional posterior distributions of $\boldsymbol{\gamma}|-$, $\boldsymbol{\beta}|-$, $\sigma^2|-$ and $\tau_j^2|-$, $j = 1, \ldots, \tilde{P}$. To this end, $\sigma^2|- \sim IG(a_\sigma + M/2, b_\sigma + ||\boldsymbol{y}_{\boldsymbol{\Phi}} - \boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\gamma}||^2/2)$, $\tau_j^2|- \sim IG(a_\tau + H/2, b_\tau + ||\boldsymbol{\gamma}_j||^2/2)$ and $\boldsymbol{\beta}|- \sim N\left(\left(\boldsymbol{X}_{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{\Phi}}/\sigma^2 + \boldsymbol{I}\right)^{-1}\boldsymbol{X}_{\boldsymbol{\Phi}}^{\mathrm{T}}(\boldsymbol{y}_{\boldsymbol{\Phi}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\gamma})/\sigma^2, \left(\boldsymbol{X}_{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{X}_{\boldsymbol{\Phi}}/\sigma^2 + \boldsymbol{I}\right)^{-1}\right)$ do not present

any computational obstacles. The main computational bottleneck lies with $\boldsymbol{\gamma}|-$,

$$N\left(\left(\frac{\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}}{\sigma^2}+\boldsymbol{\Delta}^{-1}\right)^{-1}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\frac{(\boldsymbol{y}_{\boldsymbol{\Phi}}-\boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta})}{\sigma^2},(\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}/\sigma^2+\boldsymbol{\Delta}^{-1})^{-1}\right). \quad (6)$$

Efficient sampling of $\boldsymbol{\gamma}$ uses the Cholesky decomposition of $\left(\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}/\sigma^2+\boldsymbol{\Delta}^{-1}\right)$ and solves triangular linear systems to draw a sample from (6). While numerically robust for small to moderately large H, computing and storing the Cholesky factor involves $O((H\tilde{P})^3)$ and $O((H\tilde{P})^2)$ floating point operations, respectively (Golub and Van Loan, 2012). This produces bottlenecks for a large number of basis functions, which is required to estimate the functional coefficients with sufficient local variation.

To achieve computational efficiency, we adapt a recent algorithm proposed in Bhattacharya et al. (2016) (in the context of ordinary linear regression with uncompressed data and small sample size) to our setting, with the details provided in Algorithm 1. Predictive inference on $y(\boldsymbol{u}_0)$ proceeds from the posterior predictive distribution

$$\mathbb{E}[p(y(\boldsymbol{u}_0)\,|\,\boldsymbol{y}_{\boldsymbol{\Phi}},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2)]=\int p(y(\boldsymbol{u}_0)\,|\,\boldsymbol{y}_{\boldsymbol{\Phi}},\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2)p(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2\,|\,\boldsymbol{y}_{\boldsymbol{\Phi}},\boldsymbol{\Phi})d\boldsymbol{\beta}d\boldsymbol{\gamma}d\sigma^2\,, \quad (7)$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to the posterior distribution in (4). This is easily achieved using composition sampling, as outlined in Algorithm 2.

Algorithm 1: Parametric Inference from the Proposed Model

begin
- Draw $\tilde{\boldsymbol{\gamma}}_1\sim N(\boldsymbol{0},\boldsymbol{\Delta})$ and $\tilde{\boldsymbol{\gamma}}_2\sim N(\boldsymbol{0},\boldsymbol{I}_M)$
- Set $\tilde{\boldsymbol{\gamma}}_3=\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\tilde{\boldsymbol{\gamma}}_1/\sigma+\tilde{\boldsymbol{\gamma}}_2$
- Solve $(\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}\boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}/\sigma^2+\boldsymbol{I}_M)\tilde{\boldsymbol{\gamma}}_4=((\boldsymbol{y}_{\boldsymbol{\Phi}}-\boldsymbol{X}_{\boldsymbol{\Phi}}\boldsymbol{\beta})/\sigma-\tilde{\boldsymbol{\gamma}}_3)$
- Set $\tilde{\boldsymbol{\gamma}}_5=\tilde{\boldsymbol{\gamma}}_1+\boldsymbol{\Delta}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi}}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}_4/\sigma$.
- The resulting $\tilde{\boldsymbol{\gamma}}_5$ is a draw from the full conditional posterior distribution of $\boldsymbol{\gamma}$. The computation is dominated by step (iii), which comprises $O(M^3+M^2H\tilde{P})$.
- When basis functions involve parameters, they are updated using Metropolis-Hastings steps since no closed form full conditionals are generally available for them.

end

The next section offers theoretical results on asymptotic consistency of the posterior distribution for the compressed model (4) and the posterior predictive distribution (7) with respect to the probability law for the uncompressed oracle model in (1).

3. Posterior contraction from data sketching

3.1 Definitions and Notations

This section proves the posterior contraction properties of varying coefficients under the proposed framework. In what follows, we add a subscript N to the compressed response

Algorithm 2: Predictive Inference from the Proposed Model

begin
  L denotes the number of post-convergence posterior samples. **for** $l = 1 : L$ **do**
    Draw $\{\boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}, \sigma^{2(l)}\}$ from (4)
    Draw $w_j(\boldsymbol{u}_0)^{(l)}$ from $\boldsymbol{\gamma}^{(l)}$ using (2)
    Draw $y(\boldsymbol{u}_0)^{(l)} \sim N(\sum_{p=1}^{P} x_p(\boldsymbol{u}_0)\beta_p^{(l)} + \sum_{j=1}^{\tilde{P}} \tilde{x}_j(\boldsymbol{u}_0) w_j(\boldsymbol{u}_0)^{(l)}, \sigma^{2(l)})$
  end
  $y(\boldsymbol{u}_0)^{(1)}, ..., y(\boldsymbol{u}_0)^{(L)}$ are samples from (7).
end

vector $\boldsymbol{y}_{\boldsymbol{\Phi},N}$, compressed predictor matrix $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}$, dimension of the compression matrix M_N and the number of basis functions H_N to indicate that all of them increase with the sample size N. Naturally, the dimension of the basis coefficient vector $\boldsymbol{\gamma}$ and the compression matrix $\boldsymbol{\Phi}$ are also functions of N, though we keep this dependence implicit. Since we do not assume a functional variable selection framework, we keep P fixed throughout, and not a function of N. We assume that $\boldsymbol{u}_1, ..., \boldsymbol{u}_N$ follow i.i.d. distribution G on $\mathcal{D}$ with G having a Lebesgue density g, which is bounded away from zero and infinity uniformly over $\mathcal{D}$. The true regression function is also given by (1), with the true varying coefficients $w_1^*(\boldsymbol{u}), ..., w_{\tilde{P}}^*(\boldsymbol{u})$ belonging to the class of functions

$$\mathcal{F}_\xi(\mathcal{D}) = \{f : f \in L_2(\mathcal{D}) \cap \mathcal{C}^\xi(\mathcal{D}), E_{\mathcal{U}}[|f|] < \infty\}, \tag{8}$$

where $L_2(\mathcal{D})$ is the set of all square integrable functions on $\mathcal{D}$, $\mathcal{C}^\xi(\mathcal{D})$ is the class of at least ξ-times continuously differentiable functions in $\mathcal{D}$ and $E_{\mathcal{U}}$ denotes the expectation under the density of g. The probability and expectation under the true data generating model are denoted by P^* and E^*, respectively. For algebraic simplicity, we make a few simplifying assumptions in the model. To be more specific, we assume that $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = \sigma^{*2}$ is known and fixed at 1. The first assumption is mild since P does not vary with N and we do not consider variable selection. The second assumption is also customary in asymptotic studies (Vaart and Zanten, 2011). Furthermore, the theoretical results obtained by assuming σ^2 as a fixed value is equivalent to those obtained by assigning a prior with a bounded support on σ^2 (Van der Vaart et al., 2009).

For a vector $\boldsymbol{v} = (v_1, ..., v_N)^{\mathrm{T}}$, we let $||\cdot||_1, ||\cdot||_2$ and $||\cdot||_\infty$ denote the L_1, L_2 and L_∞ norms defined as $||\boldsymbol{v}||_2 = (\sum_{n=1}^{N} v_n^2)^{1/2}$, $||\boldsymbol{v}||_1 = \sum_{n=1}^{N} |v_n|$ and $||\boldsymbol{v}||_\infty = \max_{n=1,..,N} |v_n|$, respectively. The number of nonzero elements in a vector is given by $||\cdot||_0$. In the case of a square integrable function $f(\boldsymbol{u})$ on $\mathcal{D}$, we denote the integrated L_2-norm of f by $||f||_2 = \left(\int_{\mathcal{D}} f(\boldsymbol{u})^2 g(\boldsymbol{u}) d\boldsymbol{u}\right)^{1/2}$ and the sup-norm of f by $||f||_\infty = \sup_{\boldsymbol{u}\in\mathcal{D}} |f(\boldsymbol{u})|$. Thus $||\cdot||_\infty$ and $||\cdot||_2$ are used both for vectors and functions, and they should be interpreted based on the context. Finally, $e_{\min}(\boldsymbol{A})$ and $e_{\max}(\boldsymbol{A})$ represent the minimum and maximum eigenvalues of the symmetric matrix $\boldsymbol{A}$, respectively. The Frobenius norm of the matrix $\boldsymbol{A}$ is given by $||\boldsymbol{A}||_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$. For two nonnegative sequences $\{a_N\}$ and $\{b_N\}$, we write $a_N \asymp b_N$ to denote $0 < \liminf_{N\to\infty} a_N/b_N \leq \limsup_{N\to\infty} a_N/b_N < \infty$. If $\lim_{N\to\infty} a_N/b_N = 0$, we write $a_N = o(b_N)$ or $a_N \prec b_N$. We use $a_N \lesssim b_N$ or $a_N = O(b_N)$ to denote that for sufficiently large N, there exists a constant $C > 0$ independent of N such that $a_N \leq C b_N$.

3.2 Assumption, Framework and Main Results

For simplicity, we assume that the random covariates $x_p(\boldsymbol{u})$, $p = 1, ..., P$ follow distributions which are independent of the distribution of the idiosyncratic error ϵ. We now state the following assumptions on the basis functions, H_N, M_N, covariates and the sketching or compression matrix.

(A) For any $w_j^*(\boldsymbol{u}) \in \mathcal{F}_\xi(\mathcal{D})$, there exists $\boldsymbol{\gamma}_j^*$ such that

$$||w_j^* - \boldsymbol{B}_j^{\mathrm{T}}\boldsymbol{\gamma}_j^*||_\infty = \sup_{\boldsymbol{u}\in\mathcal{D}} |w_j^*(\boldsymbol{u}) - \sum_{h=1}^{H_N} B_{jh}(\boldsymbol{u})\gamma_{jh}^*| = O(H_N^{-\xi}),$$

for $j = 1, ..., \tilde{P}$, and $||\boldsymbol{\gamma}^*||_2^2 \prec M_N^{d/(d+2\xi)}$.

(B) N, M_N, H_N satisfy $M_N = o(N)$ and $H_N \asymp M_N^{1/(2\xi+d)}$.

(C) $||\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} - \boldsymbol{I}_{M_N}||_F \leq C'\sqrt{M_N/N}$, for some constant $C' > 0$, for all large N.

(D) The random covariate $x_p(\boldsymbol{u})$ are uniformly bounded for all $\boldsymbol{u} \in \mathcal{D}$, and w.l.g., $|x_p(\boldsymbol{u})| \leq 1$, for all $p = 1, ..., P$ and for all $\boldsymbol{u} \in \mathcal{D}$.

(E) There exists a sequence κ_N such that $||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\alpha}||^2 \asymp \kappa_N||\tilde{\boldsymbol{X}}_N\boldsymbol{\alpha}||^2$, such that $1 \prec N\kappa_N \prec M_N$ for any vector $\boldsymbol{\alpha} \in \mathbb{R}^{N\tilde{P}}$.

(F) For simplicity, assume $\boldsymbol{\Delta} = \boldsymbol{I}$, $\boldsymbol{\beta} = \boldsymbol{0}$, σ^2 is known and without loss of generality, $\sigma^2 = 1$.

Assumption (A) holds for orthogonal Legendre polynomials, Fourier series, B-splines and wavelets (Shen and Ghosal, 2015). Assumption (B) provides an upper bound on the growth of M_N and H_N as a function of N. Assumption (C) is a mild assumption based on the theory of random matrices and occurs with probability at least $1 - e^{-C''M_N}$ when $\boldsymbol{\Phi}$ is constructed using the Gaussian sketching for a constant $C'' > 0$ (see Lemma 5.36 and Remark 5.40 of Vershynin (2010)). Assumption (D) is a technical condition customarily used in functional regression analysis (Bai et al., 2019). Assumption (E) characterizes the class of feasible compression matrices, roughly explaining how the linear structure of the columns of the original predictor matrix is related to that of the compressed predictor matrix. Such an assumption is reasonable for the set of random compression matrices for a sequence κ_N depending on N, M_N and $\tilde{P}$ (Ahfock et al., 2017). As argued here, both Assumptions (C) and (E) can be proved to hold with high probability. We include them as assumptions because they are considered to hold with probability 1. This practice is common when random matrices are used in the study of computationally efficient Bayesian models (Guhaniyogi and Dunson, 2015, 2016; Guhaniyogi and Scheffler, 2021), as it allows the focus to remain on model uncertainty without factoring in the uncertainty of random matrix construction. Finally, Assumption (F) is assumed for simplicity in mathematical derivation, and it could potentially be relaxed.

Let $\boldsymbol{w}(\boldsymbol{u}) = (w_1(\boldsymbol{u}), ..., w_{\tilde{P}}(\boldsymbol{u}))^{\mathrm{T}}$ and $\boldsymbol{w}^*(\boldsymbol{u}) = (w_1^*(\boldsymbol{u}), ..., w_{\tilde{P}}^*(\boldsymbol{u}))^{\mathrm{T}}$ be the $\tilde{P}$-dimensional fitted and true varying coefficients. Let $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2 = \sum_{j=1}^{\tilde{P}} \|w_j - w_j^*\|_2$ denote the sum

of integrated L_2 distances between the true and the fitted varying coefficients. Define the set $\mathcal{C}_N = \left\{\boldsymbol{w} : ||\boldsymbol{w} - \boldsymbol{w}^*||_2 > \tilde{C}\theta_N\right\}$, for some constant $\tilde{C}$ and some sequence $\theta_N \to 0$ and $M_N\theta_N^2 \to \infty$. In addition, suppose that $\pi_N(\cdot)$ and $\Pi_N(\cdot)$ are the prior and posterior densities of $\boldsymbol{w}$ with N observations, respectively. From equation (2), the prior distribution on $\boldsymbol{w}$ is governed by the prior distribution on $\boldsymbol{\gamma}$, so that the posterior probability of $\mathcal{C}_N$ is

$$\Pi_N(\mathcal{C}_N|\boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}) = \frac{\int_{\mathcal{C}_N} f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{\gamma})\pi_N(\boldsymbol{\gamma})}{\int f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{\gamma})\pi_N(\boldsymbol{\gamma})},$$

where $f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{\gamma})$ is the joint density of $\boldsymbol{y}_{\boldsymbol{\Phi},N}$ under model (4). We begin with the following important result from the random matrix theory.

Lemma 1 *Consider the $M_N \times N$ compression matrix $\boldsymbol{\Phi}$ with each entry drawn independently from $N(0, 1/N)$. Also, assume that $M_N = o(N)$. Then, almost surely*

$$(\sqrt{N} - o(\sqrt{N}))^2/N \leq e_{\min}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}) \leq e_{\max}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}) \leq (\sqrt{N} + o(\sqrt{N}))^2/N, \tag{9}$$

when $N \to \infty$.

Proof This is a consequence of Theorem 5.31 and Corollary 5.35 of Vershynin (2010). ■

The inequalities in (9) are used to derive the following two results, which we present as Lemma 2 and 3.

Lemma 2 *Let P^* denote the true probability distribution of $\boldsymbol{y}_N$ and $f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)$ denotes the density of $\boldsymbol{y}_{\boldsymbol{\Phi},N}$ (omitting explicit dependence on $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}$) under the true data generating model. Define*

$$\mathcal{A}_N = \left\{\boldsymbol{y} : \int \left\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\right\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma} \leq \exp(-CM_N\theta_N^2)\right\}. \tag{10}$$

Then $P^*(\mathcal{A}_N) \to 0$ as $M_N, N \to \infty$ for any constant $C > 0$.

Proof See Section A in the Appendix. ■

Lemma 3 *Let $\boldsymbol{\gamma}^*$ be any fixed vector in the support of $\boldsymbol{\gamma}$ and let $\mathcal{B}_N = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 \leq C_{2w}\theta_N H_N^{1/2}\}$ for some constant $C_{2w} > 0$. Then there exists a sequence ζ_N of random variables depending on $\{\boldsymbol{y}_{\boldsymbol{\Phi},N}, \boldsymbol{X}_{\boldsymbol{\Phi},N}\}$ and taking values in $(0, 1)$ such that*

$$\mathbb{E}^*(\zeta_N) \lesssim \exp(-M_N\theta_N^2) \text{ and } \sup_{\boldsymbol{\gamma} \in \mathcal{B}_N^c} \mathbb{E}_{\boldsymbol{\gamma}}(1 - \zeta_N) \lesssim \exp(-M_N\theta_N^2), \tag{11}$$

where $\mathbb{E}_{\boldsymbol{\gamma}}$ and $\mathbb{E}^*$ denote the expectations under the distributions $f(\cdot\,|\,\boldsymbol{\gamma})$ and $f^*(\cdot\,|\,\boldsymbol{\gamma}^*)$, respectively.

Proof See Section B in the Appendix. ■

We use the above results to establish the posterior contraction result for the proposed model.

Theorem 4 *Under Assumptions (A)-(F), our proposed model (4) satisfies*

$$\max_{j=1,\ldots,\tilde{P}} \sup_{w_j^* \in \mathcal{F}_\xi(\mathcal{D})} \mathbb{E}^* \Pi_N(\mathcal{C}_N \mid \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}) \to 0, \text{ as } N, M_N \to \infty,$$

with the posterior contraction rate $\theta_N \asymp M_N^{-\xi/(2\xi+d)}$.

Proof See Section C in the Appendix for the detailed proof. Here we offer an outline of the proof. The steps are given below.
Step 1: Using basis expansion of each w_j and by Assumption (A), $\{\boldsymbol{w} : ||\boldsymbol{w} - \boldsymbol{w}^*||_2 \geq \tilde{C}\theta_N\} \subset \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 H_N^{-1/2} \geq C_{2w}\theta_N\} = \mathcal{B}_N^c$, for some constant $C_{2w} > 0$.
Step 2: Consider the set $\mathcal{A}_N$ defined in Lemma 2 and the sequence of random variables ζ_N defined in Lemma 3. Note that $\mathbb{E}^*(\zeta_N) \to 0$ as $N \to \infty$, by Lemma 3 and $P^*(\mathcal{A}_N) \to 0$ as $N \to \infty$, by Lemma 2.
Step 3: We then consider the expression $\mathbb{E}^*[\Pi(\mathcal{B}_N^c \mid \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N})(1-\zeta_N)1_{y_N \in \mathcal{A}_N^c}]$
$= \mathbb{E}^* \left[1_{\boldsymbol{y}_N \in \mathcal{A}_N^c} \frac{\left\{(1-\zeta_N)\int_{\mathcal{B}_N^c} \{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right\}}{\left\{\int \{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right\}} \right]$. Due to Step 2, it only suffices to show that this expression converges to 0 as $N \to \infty$.
Step 4: Under Assumptions (A)-(F), the numerator of the above expression decays exponentially to 0 as a function of $M_N\theta_N^2$. The inverse of the denominator grows at a slower rate of $M_N\theta_N^2$. The result is then proved by considering $M_N\theta_N^2 = o(N)$. ■

Since $\theta_N \to 0$ as $N \to \infty$, the model consistently estimates the true varying coefficients under the integrated L_2-norm. Further, data compression decreases the effective sample size from N to M_N, hence, the contraction rate θ_N obtained in Theorem 4 is optimal and adaptive to the smoothness of the true varying coefficients. Our next theorem justifies the two-stage prediction strategy described in Section 2.2.

Theorem 5 *For any input* $\boldsymbol{u}_0$ *drawn randomly with the density* g *and corresponding predictors* $\tilde{x}_1(\boldsymbol{u}_0), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}_0)$*, let* f_u *be the predictive density* $p(y(\boldsymbol{u}_0) \mid \tilde{x}_1(\boldsymbol{u}_0), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}_0), w(\boldsymbol{u}_0))$ *derived from (1) without data compression. Let* f^* *be the true data generating model (i.e., (1) with* $\boldsymbol{w}(\boldsymbol{u}_0)$ *fixed at* $\boldsymbol{w}^*(\boldsymbol{u}_0)$*). Given* $\boldsymbol{u}_0$ *and* $\tilde{x}_1(\boldsymbol{u}_0), \ldots, \tilde{x}_{\tilde{P}}(\boldsymbol{u}_0)$*, define* $h(f_u, f^*) = \int(\sqrt{f_u} - \sqrt{f^*})^2$ *as the Hellinger distance between the densities* f_u *and* f^**. Then*

$$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{U}}[h(f_u, f^*) \mid \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{y}_{\boldsymbol{\Phi},N}] \to 0, \text{ as } N, M_N \to \infty, \tag{12}$$

where $\mathbb{E}_{\mathcal{U}}$*,* $\mathbb{E}$ *and* $\mathbb{E}^*$ *stand for expectations with respect to the density* g*, the posterior density* $\Pi_N(\cdot|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}, \boldsymbol{y}_{\boldsymbol{\Phi},N})$ *and the true data generating distribution, respectively.*

Proof See Section D in the Appendix. ■

The theorem states that the predictive density of the VCM model in (1) is arbitrarily close to the true predictive density even when we plug-in inference on parameters from (4).

4. Simulation Results

4.1 Inferential performance

We empirically validate our proposed approach using (4) for $d = 2$, i.e., for the spatially varying coefficient models. The approach, henceforth abbreviated as *geoS*, is compared with the uncompressed model (3) on some simulated data in terms of inferential performance and computational efficiency. We simulate data by using a fixed set of spatial locations $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ that were drawn uniformly over the domain $\mathcal{D} = [0,1] \times [0,1]$. We set $\tilde{P} = P = 3$ and assume $\boldsymbol{\beta} = 0$, i.e., all predictors have purely space-varying coefficients. We set $\tilde{x}_1(\boldsymbol{u}_i) = 1$, for all $i = 1, \ldots, N$, while the values of $\tilde{x}_j(\boldsymbol{u}_1), \ldots, \tilde{x}_j(\boldsymbol{u}_N)$ for $j = 2, 3$ were set to independently values from $N(0,1)$. For each $n = 1, \ldots, N$, the response $y(\boldsymbol{u}_n)$ is drawn independently from $N(w_1^*(\boldsymbol{u}_n) + w_2^*(\boldsymbol{u}_n)\tilde{x}_2(\boldsymbol{u}_n) + w_3^*(\boldsymbol{u}_n)\tilde{x}_3(\boldsymbol{u}_n), \sigma^{*2})$ following (3), where σ^{*2} is set to be 0.1. The true space-varying coefficients ($w_j^*(\boldsymbol{u})$s) are simulated from a Gaussian process with mean 0 and covariance kernel $C(\cdot, \cdot; \theta_j)$, i.e., $(w_j^*(\boldsymbol{u}_1), ..., w_j^*(\boldsymbol{u}_N))^{\mathrm{T}}$ is drawn from $N(0, C^*(\theta_j))$, for each $j = 1, \ldots, \tilde{P}$, where $C^*(\theta_j)$ is an $N \times N$ matrix with the (n, n')th element $C(\boldsymbol{u}_n, \boldsymbol{u}_{n'}; \theta_j)$. We set the covariance kernel $C(\cdot, \cdot; \theta_j)$ to be the exponential covariance function given by

$$C(\boldsymbol{u}, \boldsymbol{u}'; \theta_j) = \delta_j^2 \exp\left\{-\frac{1}{2}\left(\frac{||\boldsymbol{u} - \boldsymbol{u}'||}{\phi_j}\right)\right\}, \;\; j = 1, 2, 3, \tag{13}$$

with the true values of $\delta_1^2, \delta_2^2, \delta_3^2$ set to $1, 0.8, 1.1$, respectively. We fix the true values of ϕ_1, ϕ_2, ϕ_3 at $1, 1.25, 2$, respectively.

While fitting *geoS* and its uncompressed analogue (3), the varying coefficients are modeled through the linear combination of H basis functions as in (2), where these basis functions are chosen as the tensor-product of B-spline bases of order $q = 4$ (Shen and Ghosal, 2015). More specifically, for $\boldsymbol{u} = (u^{(1)}, u^{(2)})$, the j-th varying coefficient is modeled as

$$w_j(\boldsymbol{u}) = \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} B_{jh_1}^{(1)}(u^{(1)}) B_{jh_2}^{(2)}(u^{(2)}) \gamma_{jh_1h_2} \;, \tag{14}$$

where the marginal B-splines $B_{jh_1}^{(1)}$, $B_{jh_2}^{(2)}$ are defined on sets of H_1 and H_2 knots, respectively. The knots are chosen to be equally-spaced so the entire set of $H = H_1H_2$ knots is uniformly spaced over the domain $\mathcal{D}$. We complete the hierarchical specification by assigning independent $IG(2, 0.1)$ priors (mean 0.1 with infinite variance) for σ^2 and τ_j^2 for each $j = 1, \ldots, P$.

We implemented our models in the `R` statistical computing environment on a Dell XPS 13 PC with Intel Core i7-8550U CPU @ 4.00GHz processors at 16 GB of RAM. For each of our simulation data sets we ran a single-threaded MCMC chain for 5000 iterations. Posterior inference was based upon 2000 samples retained after adequate convergence was diagnosed using Monte Carlo standard errors and effective sample sizes (ESS) using the `mcmcse` package in `R`. Source codes for these experiments are available from `https://github.com/LauraBaracaldo/Bayesian-Data-Sketching-in-Spatial-Regression-Models`.

Table 1 summarizes the estimates of the varying coefficients and the predictive performance for *geoS* in comparison to the uncompressed model. These results are based on K

	$N = 5000$, $H = 225$, $K = 50$	
	(geoS) $M = 700$	*Uncompressed*
MSE (SVC)	0.0335 (0.028, 0.039)	0.0109
95% CI length	0.6751 (0.654, 0.723)	0.2441
95% CI coverage	0.9406 (0.913, 0.959)	0.9520
MSPE	0.1986 (0.174, 0.231)	0.1369
95% PI length	1.6449 (1.567, 1.755)	1.3071
95% PI coverage	0.9400 (0.912, 0.962)	0.9380
Computation efficiency	2.1165 (1.643, 2.152)	0.6298

	$N = 10000$, $H = 256$, $K = 50$	
	(geoS) $M = 1000$	*Uncompressed*
MSE (SVC)	0.0238 (0.019, 0.028)	0.0092
95% CI length	0.6101 (0.591, 0.631)	0.2837
95% CI coverage	0.9253 (0.920, 0.960)	0.9500
MSPE	0.1737 (0.156, 0.191)	0.1260
95% PI length	1.6013 (1.534, 1.653)	1.3770
95% PI coverage	0.9460 (0.928, 0.965)	0.9510
Computation efficiency	2.2368 (2.101, 2.288)	0.4981

	$N = 100000$, $H = 400$, $K = 10$	
	(geoS) $M = 3200$	*Uncompressed*
MSE (SVC)	0.0067 (0.003, 0.008)	0.0008
95% CI length	0.3007 (0.221, 0.310)	0.1712
95% CI coverage	0.9360 (0.926, 0.941)	0.953
MSPE	0.1242 (0.115, 0.131)	0.112
95% PI length	1.3503 (1.290, 1.381)	1.239
95% PI coverage	0.9510 (0.942, 0.956)	0.937
Computation efficiency	5.9081 (5.814, 6.001)	0.981

Table 1: Summary results: 50% (2.5%, 97.5%) over K simulations for the compressed *geoS* model. Median values for each metric over K simulations are presented for the uncompressed model. Mean Squared Error (MSE), length and coverage of 95% CI for the spatially varying coefficients are presented. We also provide mean squared prediction error (MSPE), coverage and length of 95% predictive intervals for competing models.

independently generated data sets for each scenario, with $N = 5,000$ (case 1), $N = 10,000$ (case 2), and $N = 100,000$ (case 3). For each case, the compressed dimension is taken to be $M \approx 10\sqrt{N}$, which seems to be effective from empirical considerations in our simulations. We provide further empirical justification for this choice in Section 4.2. Our *geoS* approach compresses the sample sizes to $M = 700$, $M = 1000$ and $M = 3000$ in cases 1, 2 and 3, respectively. The number of fitted basis functions in cases 1, 2 and 3 are $H = 225, 256, 400$, respectively. For each simulated data, we evaluated 6 model assessment metrics and 1 computational efficiency metric that are listed in Table 1. We present the median, 2.5, and 97.5 quantiles for each of the metrics on the K data sets. We see that *geoS* offers competitive inferential performance and outperforms the uncompressed model. For example, the confidence interval for some of the metrics, while including the values for the uncompressed model, often reveal heavier mass to the left of the value for the uncompressed model.

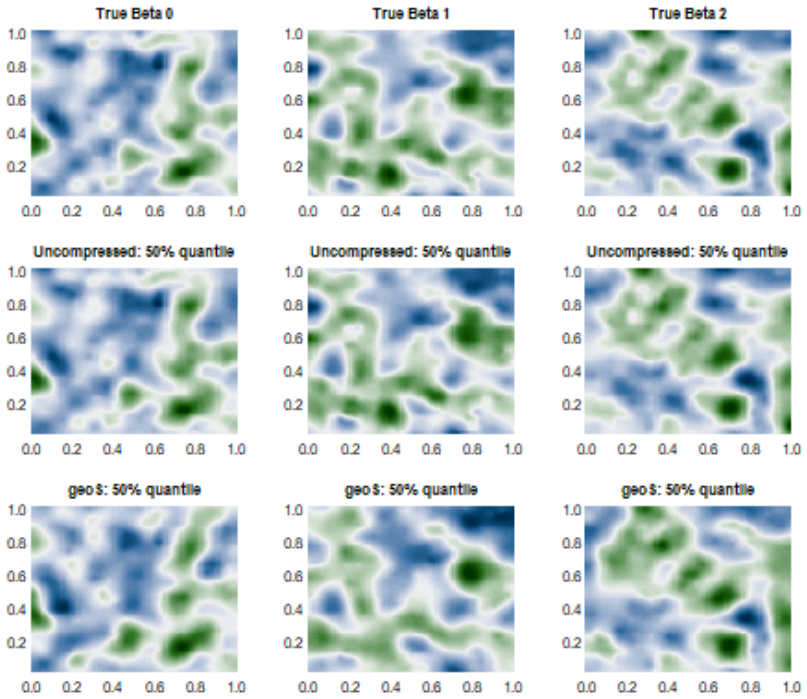


Figure 1: Simulation case 1: $(N, H) = (5000, 225)$. Two-dimensional true and predicted surfaces over the unit square $\mathcal{D} = [0,1] \times [0,1]$. First row corresponds to the surfaces of true space-varying coefficients $\beta_p^*(s)$, $p = 1, 2, 3$. Rows 2 and 3 correspond to the predicted 50% quantile surfaces for the uncompressed and compressed *geoS* models respectively.

Figures 1 and 2 present the estimated varying coefficients in one representative simulation experiment by *geoS* and the uncompressed data model for cases 1 and 2, respectively. These figures reveal point estimates that are substantively similar to those from *geoS* and the uncompressed model. The mean squared error of estimating varying coefficients, defined as $\sum_{j=1}^{3}\sum_{n=1}^{N}(\widehat{w}_j(\boldsymbol{u}_n) - w_j^*(\boldsymbol{u}_n))^2/(3N)$ (where $\widehat{w}_j(\boldsymbol{u}_n)$ is the posterior median of $w_j(\boldsymbol{u}_n)$), also confirms very similar point estimates offered by the compressed and uncompressed models (see Table 1). Further, *geoS* offers close to nominal coverage for 95% credible intervals for varying coefficients, with little wider credible intervals compared to uncompressed data model. This can be explained by the smaller sample size for the *geoS* model, though the difference turns out to be minimal. We also carry out predictive inference using *geoS* (Section 2.2). Table 1 presents mean squared predictive error (MSPE), average length and coverage for the 95% predictive intervals, based on $N^* = 500$ out of the sample observations. We find *geoS* delivers posterior predictive estimates and predictive coverage that are very consistent with the uncompressed model, perhaps with marginally wider predictive intervals than those without compression. Finally, the computational efficiency of both models are computed based on the metric $\log_2(ESS/\text{Computation Time})$, where ESS denotes the effective sample size averaged over the MCMC samples of all parameters. We find *geoS* is almost 240%, 350% and 500% more efficient than the uncompressed model

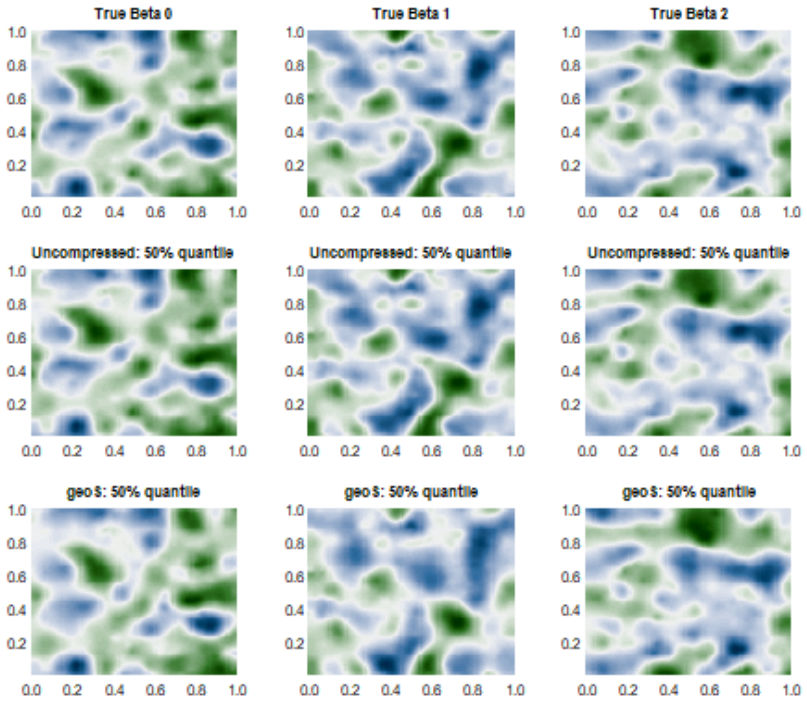


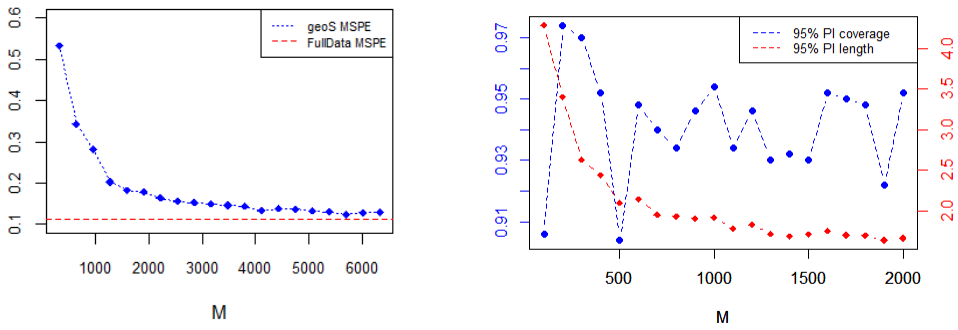
Figure 2: Simulation case 1: $(N, H) = (10000, 256)$. Two-dimensional true and predicted surfaces over the unit square $\mathcal{D} = [0, 1] \times [0, 1]$. First row corresponds to the surfaces of true space-varying coefficients $\beta_p^*(s)$, $p = 1, 2, 3$. Rows 2 and 3 correspond to the predicted 50% quantile surfaces for the uncompressed and compressed *geoS* models respectively.

for $N = 5,000$, $N = 10,000$ and $N = 100,000$, respectively, while delivering substantively consistent inference on the spatial effects.

We conduct an additional experiment where we compare our method with a sparse Gaussian Process (GP) model. For each $n = 1, \ldots$ the simulated response $y(\boldsymbol{u}_n)$ is drawn independently from $N(w^*(\boldsymbol{u}_n), \sigma^{*2})$. Table 2 summarizes results in terms of inferential performance and computational efficiency for *geoS* compared to the *predictive process model* (e.g., Banerjee et al., 2008). The predictive process achieves a reduction in computational complexity by projecting the original Gaussian Process (GP) onto a lower-dimensional subspace defined by a set of knots or inducing points, and is envisioned as a sparse GP model. It is implemented using the `spBayes` package in `R`. Notably, the `spBayes` package only allows fitting a varying intercept model with GP or predictive process fitted on the varying intercept, which prompted us to simulate the data as above. Our findings indicate that our method and the predictive process exhibit very similar inferential performance in terms of accuracy and predictive capability. However, *geoS* demonstrates superior computational efficiency across all evaluated scenarios. Unlike the sparse GP, which experiences increased computational demands with larger M, *geoS* maintains a consistent computational profile in terms of scalability and offers a more practical method for handling very large datasets.

		GeoS	*SparseGP*
$M = 710$	*MSPE*	0.110	0.103
	95% PI length	1.320	1.503
	95% PI coverage	0.940	0.942
	Time (secs)	17.47	5190.77
$M = 337$	*MSPE*	0.120	0.111
	95% PI length	1.396	1.316
	95% PI coverage	0.932	0.944
	Time (secs)	14.23	1217.64
$M = 142$	*MSPE*	0.146	0.128
	95% PI length	1.561	1.414
	95% PI coverage	0.918	0.931
	Time (secs)	12.121	242.67

Table 2: Performance comparison of GeoS and Predictive Process for different values of $M = k\sqrt{N}$, for $k = 2, 5, 10$ and $N = 5000$



(a) (b)

Figure 3: (a) MSPE, (b) 95% predictive interval coverage and length for different choices of M

4.2 Choice of the dimension of the compression matrix M

We present investigations into the appropriate compression matrix size M. For simulated data with sample size $N = 100000$, we ran our model for different values of $M = k\sqrt{N}$, $k = 1, \ldots, 20$. Figure 3 shows the variations in the prediction of points and intervals reflected in the $MSPE$ and 95% predicted interval coverage and length, respectively. Unsurprisingly, as M increases the MSPE drops with a decreased rate of decline until $k \sim 10$. In terms of interval prediction, the predictive coverage seems to oscillate within the narrow interval $(0.9, 0.97)$ for all values of M, but the length of the predictive interval improves as M increases and begins to stabilize around $k \sim 10$. We observe that the choice of $M \sim 10\sqrt{N}$ leads to good performance in various simulations and real data analysis.

Table 3: Median and 95% credible interval of β_1, β_2 for geoS and its uncompressed analogue are presented for the Vegetation data analysis. We also present MSPE, coverage and length of 95% predictive intervals for the competing models. Computational efficiency for the two competing models are also provided.

	(geoS) $M = 2300$	*Uncompressed*
β_1	0.222 (0.212, 0.230)	0.229 (0.219, 0.237)
β_2	-0.060 (-0.074, -0.047)	-0.071 (-0.082, -0.059)
$MSPE$	0.00327	0.00276
95% PI length	0.23445	0.22136
95% PI coverage	0.95250	0.95411
Computation efficiency	3.5424	0.46901

5. Vegetation Data Analysis

We implement *geoS* to analyze vegetation data gathered through the Moderate Resolution Imaging Spectroradiometer (MODIS), which resides aboard the Terra and Aqua platforms on NASA spacecrafts. MODIS vegetation indices, produced on 16-day intervals and at multiple spatial resolutions, provide consistent information on the spatial distribution of vegetation canopy greenness, a composite property of leaf area, chlorophyll and canopy structure. The variable of interest will be the Normalized Difference Vegetation Index (NDVI), which quantifies the relative vegetation density for each pixel in a satellite image, by measuring the difference between the reflection in the near-infrared spectrum (NIR) and the red light reflection (RED): $NDVI = \frac{NIR-RED}{NIR+RED}$. High NDVI values, ranging between 0.6 and 0.9 indicate high density of green leaves and healthy vegetation, whereas low values, 0.1 or below, correspond to low or absence of vegetation as in the case of urbanized areas. When analyzed over different locations, NDVI can reveal changes in vegetation due to human activities such as deforestation and natural phenomena such as wild fires and floods.

We analyze geographical data mapped on a projected sinusoidal grid (SIN), located on the western coast of the United States, more precisely zone *h08v05*, between $30^\circ N$ to $40^\circ N$ latitude and $104^\circ W$ to $130^\circ W$ longitude (see Figure 4(a)). The data, which were downloaded using the `R` package MODIS, comprises $133,000$ observed locations where the response was measured using the MODIS tool over a 16-day period in April 2016.. We retained $N = 113,000$ observations (randomly chosen) for model fitting and used the rest for prediction. In order to fit (1), we set $y(\boldsymbol{u}_n)$ to be the transformed NDVI ($\log(NDVI)+1$), $P = \tilde{P} = 2$ and consider the $P \times 1$ vector of predictors that includes an intercept and a binary index of urban area, both with fixed effects and spatially varying coefficients, i.e., $\boldsymbol{x}(\boldsymbol{u}_n) = \tilde{\boldsymbol{x}}(\boldsymbol{u}_n) = (1,\ x_2(\boldsymbol{u}_n))^{\mathrm{T}}$, with $x_2(\boldsymbol{u}_n) = \mathbb{1}_U(\boldsymbol{u}_n)$, where U denotes an urban area.

As in Section 4, we fit *geoS* with $M \sim 10\sqrt{N} = 2300$ and its uncompressed counterpart (3), by modeling the varying coefficients through a linear combination of basis functions constructed using the tensor product of B-splines of order $q = 4$ as in (14). We set $H = H_1 H_2 = 39^2 = 1521$ uniformly distributed knots in the domain $\mathcal{D}$, which results in $HP = 3042$ basis coefficients γ_{jh} that are estimated. The specification of the priors are identical to the simulation studies for σ^2, and τ_j^2, while β_j is assigned a flat prior for $j = 1, \ldots, P$.

We ran an MCMC chain for 5000 iterations and retained 2000 samples for posterior inference after adequate convergence was diagnosed. The posterior mean of β_1 and β_2, along with their estimated 95% credible intervals corresponding to *geoS* and the uncompressed model are presented in Table 3. Additionally, Table 3 offers predictive inference from both competitors based on $N^* = 20,000$ test observations. According to both models there is a global pattern of relatively low vegetation density for areas with positive urban index as the estimated slope coefficient β_2 is negative in the compressed *geoS* and in the uncompressed models. In terms of point prediction and quantification of predictive uncertainty, the two competitors offer practically indistinguishable results, as revealed by Table 3.

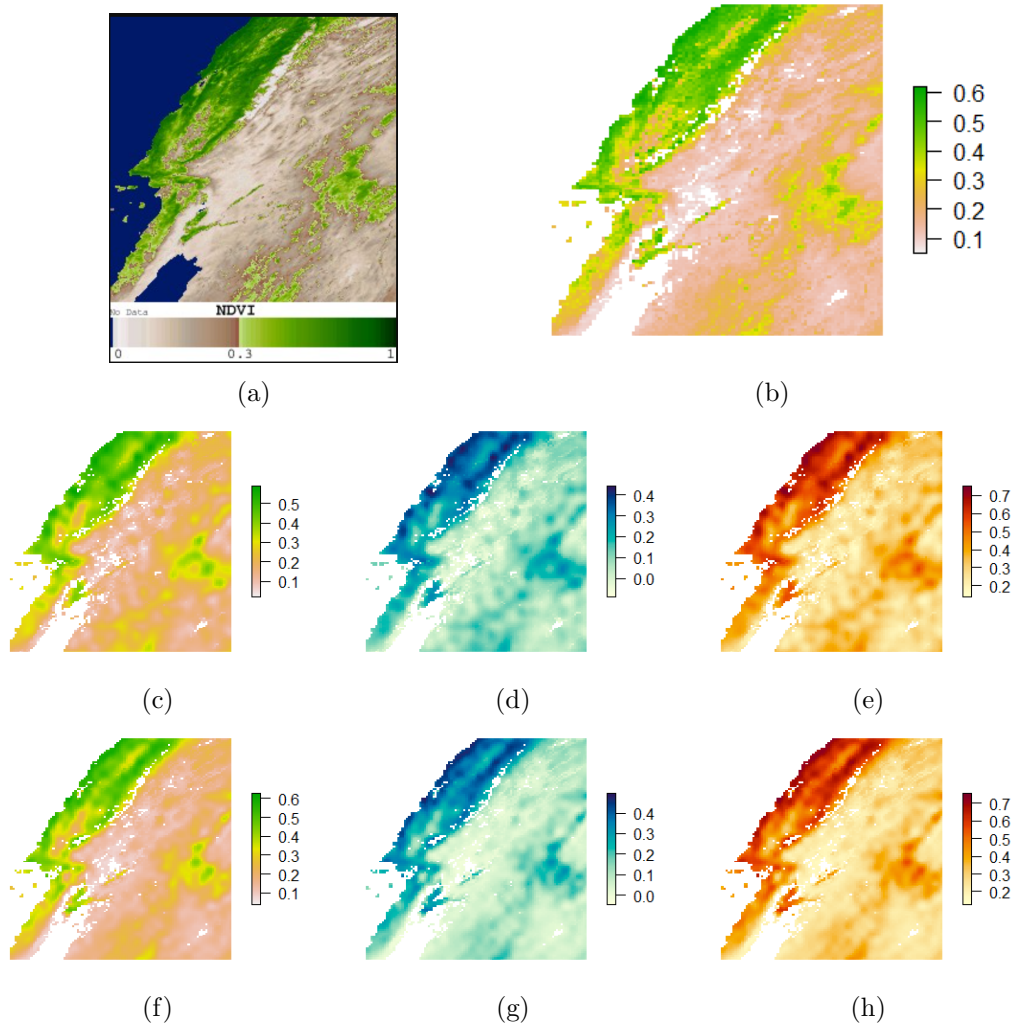


Figure 4: Colored NDVI images of western United States (zone h08v05). (a) Satellite image: MODIS/Terra Vegetation Indices 16-Day L3 Global 1 km SIN Grid - 2016.04.06 to 2016.04.21; (b) True NDVI surface (raw data). Figures (c), (d) & (e) present NVDI predicted 50%, 2.5% and 97.5% quantiles for the *geoS* model. Figures (f), (g) & (h) present NVDI Predicted 50%, 2.5% and 97.5% quantiles for the uncompressed model.

Further, Figure 4 shows that the 2.5%, 50% and 97.5% quantiles for the posterior predictive distribution are almost identical for the two competitors across the spatial domain, with the exception of neighborhoods around locations having lower NDVI values. Notably, *geoS* offers nominal coverage for 95% prediction intervals, even with a significant reduction in the sample size from $N = 113,000$ to $M = 2300$. Data sketching to such a scale considerably reduces the computation time, leading to a much higher computation efficiency of *geoS* in comparison with its uncompressed analogue.

6. Summary

We have developed Bayesian sketching for functional response and predictor variables using varying coefficient regression models. The method achieves dimension reduction by compressing the data using a random linear transformation. The approach is different from the prevalent methods for large functional data in that no new models or algorithms need to be developed since those available for existing varying coefficient regression models can be directly applied to the compressed data. We establish attractive concentration properties of the posterior and posterior predictive distributions and empirically demonstrate the effectiveness of this method for analyzing large functional data sets.

Acknowledgments

Rajarshi Guhaniyogi acknowledges funding from the National Science Foundation through DMS-2220840 and DMS-2210672; funding from the Office of Naval Research through N00014-18-1-274; and funding from the National Institute of Neurological Disorders and Stroke (NIH/NINDS) through R01NS131604. Sudipto Banerjee acknowledges funding from the National Science Foundation through DMS-1916349 and DMS-2113778; and funding from the National Institute of Health through NIEHS-R01ES027027 and NIEHS-R01ES030210.

Appendix A. Proof of Lemma 2

Proof Define

$$\mathcal{A}_{1N} = \left\{K(f^*, f) \leq M_N\theta_N^2,\ V(f^*, f) \leq M_N\theta_N^2\right\}. \tag{15}$$

By Lemma 10 in Ghosal et al. (2007), to show (10) it is enough to show that $\Pi(\mathcal{A}_{1N}) \gtrsim \exp(-C_2 M_N\theta_N^2)$ for some constant $C_2 > 0$. Let e_k, $1 \leq k \leq M_N$ be the ordered eigenvalues of $(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}})^{-1}$. After some calculations, we derive the following expressions,

$$K(f^*, f) = \frac{1}{2}\left\{\sum_{k=1}^{M_N}(e_k - 1 - \log(e_k)) + \mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2\right]\right\} \text{ and}$$

$$V(f^*, f) = \sum_{k=1}^{M_N}\frac{(1-e_k)^2}{2} + \mathbb{E}_{\mathcal{U}}E_{\mathcal{X}}\left[||(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}})^{-1}(\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*) - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*)||_2^2\right], \tag{16}$$

where $\boldsymbol{\eta}^* = (\boldsymbol{\eta}^*(\boldsymbol{u}_1)^{\mathrm{T}}, ..., \boldsymbol{\eta}^*(\boldsymbol{u}_N)^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\eta}^*(\boldsymbol{u}) = (\eta_1^*(\boldsymbol{u}), \ldots, \eta_{\tilde{P}}^*(\boldsymbol{u}))^{\mathrm{T}}$, and $\eta_j^*(\boldsymbol{u}) = w_j^*(\boldsymbol{u}) - \sum_{h=1}^{H_N} B_{jh}(\boldsymbol{u})\gamma_{jh}^*$. Expanding $\log(e_k)$ in the powers of $(1 - e_k)$ and using Lemma 1 in Jeong

and Ghosal (2020) we find $(e_k - 1 - \log(e_k)) \sim (1-e_k)^2/2$. Another use of Lemma 1 in Jeong and Ghosal (2020) yields $\sum_{k=1}^{M_N}(1-e_k)^2 \lesssim ||\boldsymbol{I} - \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}||_F^2 \lesssim M_N/N \leq M_N\theta_N^2$. Using Lemma 1, $e_k \asymp 1$ for all $k = 1, ..., M_N$. Hence, from (16)

$$\begin{aligned}\Pi(\mathcal{A}_{1N}) &\gtrsim \Pi\left(\left\{\boldsymbol{\gamma} : \mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*) - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2\right] \lesssim M_N\theta_N^2\right\}\right) \\ &\geq \Pi\left(\left\{\boldsymbol{\gamma} : \mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)||_2^2\right] + \mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2\right] \lesssim M_N\theta_N^2/2\right\}\right),\end{aligned} \tag{17}$$

where we use $||\boldsymbol{a}-\boldsymbol{b}||_2^2 \leq 2(||\boldsymbol{a}||_2^2 + ||\boldsymbol{b}||_2^2)$, for all $\boldsymbol{a}, \boldsymbol{b}$. Let $\boldsymbol{B}_j(\boldsymbol{u}_n) = (B_{j1}(\boldsymbol{u}_n), ..., B_{jH_N}(\boldsymbol{u}_n))^{\mathrm{T}}$, for $n = 1, ..., N$ and $j = 1, ..., \tilde{P}$. By Assumption (E),

$$\begin{aligned}\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)||_2^2\right] &\asymp \kappa_N E_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_N\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)||_2^2\right] \\ &= \kappa_N(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)^{\mathrm{T}}\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_N^{\mathrm{T}}\tilde{\boldsymbol{X}}_N\boldsymbol{B}\right](\boldsymbol{\gamma}-\boldsymbol{\gamma}^*).\end{aligned}$$

Recalling that $\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_N^{\mathrm{T}}\tilde{\boldsymbol{X}}_N\boldsymbol{B}$ is a $H_N\tilde{P} \times H_N\tilde{P}$ matrix with the (j,j')-th block given by $\sum_{n=1}^N \tilde{x}_j(\boldsymbol{u}_n)\boldsymbol{B}_j(\boldsymbol{u}_n)\boldsymbol{B}_{j'}(\boldsymbol{u}_n)^{\mathrm{T}}\tilde{x}_{j'}(\boldsymbol{u}_n)$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{U}}E_{\mathcal{X}}\left[\sum_{n=1}^N \tilde{x}_j(\boldsymbol{u}_n)\boldsymbol{B}_j(\boldsymbol{u}_n)\boldsymbol{B}_{j'}(\boldsymbol{u}_n)^{\mathrm{T}}\tilde{x}_{j'}(\boldsymbol{u}_n)\right] &\asymp \mathbb{E}_{\mathcal{U}}\left[\sum_{n=1}^N \boldsymbol{B}_j(\boldsymbol{u}_n)\boldsymbol{B}_{j'}(\boldsymbol{u}_n)^{\mathrm{T}}\right] \\ &= N\mathbb{E}_{\mathcal{U}}\left[\boldsymbol{B}_j(\boldsymbol{u}_1)\boldsymbol{B}_{j'}(\boldsymbol{u}_1)^{\mathrm{T}}\right],\end{aligned}$$

where the last equation follows since $\boldsymbol{u}_1, ..., \boldsymbol{u}_N$ are i.i.d.. Hence,

$$\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)||_2^2\right] \asymp N\kappa_N\mathbb{E}_{\mathcal{U}}\left[||\boldsymbol{B}(\boldsymbol{u}_1)(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)||_2^2\right] \asymp N\kappa_N||\boldsymbol{\gamma}-\boldsymbol{\gamma}^*||_2^2/H_N, \tag{18}$$

where $\boldsymbol{B}(\boldsymbol{u}) = [\boldsymbol{B}_1(\boldsymbol{u}) : \cdots : \boldsymbol{B}_{\tilde{P}}(\boldsymbol{u})]^{\mathrm{T}}$. The last expression follows from Lemma A.1 of Huang et al. (2004). From Assumption (E) again,

$$\begin{aligned}\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2\right] \asymp \kappa_N\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{\boldsymbol{X}}_N\boldsymbol{\eta}^*||_2^2\right] &= \kappa_N\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[\sum_{n=1}^N\sum_{j=1}^{\tilde{P}}\tilde{x}_j(\boldsymbol{u}_n)^2\eta_j^*(\boldsymbol{u}_n)^2\right] \\ &\asymp \kappa_N\mathbb{E}_{\mathcal{U}}\left[\sum_{n=1}^N\sum_{j=1}^{\tilde{P}}\eta_j^*(\boldsymbol{u}_n)^2\right] \lesssim N\kappa_N H_N^{-2\xi},\end{aligned} \tag{19}$$

where the last inequality follows from Assumption (A). From (17),

$$\begin{aligned}\Pi(\mathcal{A}_{1N}) &\gtrsim \Pi\left(\boldsymbol{\gamma} : N\kappa_N||\boldsymbol{\gamma}-\boldsymbol{\gamma}^*||_2^2/H_N + N\kappa_N H_N^{-2\xi} \lesssim M_N\theta_N^2/2\right) \\ &\gtrsim \Pi\left(\boldsymbol{\gamma} : N\kappa_N||\boldsymbol{\gamma}-\boldsymbol{\gamma}^*||_2^2 \leq M_N H_N\theta_N^2\right),\end{aligned}$$

where the last step follows from Assumptions (B) and (E). Using the fact that $\int_a^b \exp(-x^2/2)dx \geq \exp(-(a^2+b^2)/2)(b-a)$, we obtain

$$\Pi\left(\boldsymbol{\gamma} : N\kappa_N||\boldsymbol{\gamma}-\boldsymbol{\gamma}^*||_2^2 \leq M_N H_N \theta_N^2\right) \geq \prod_{h,j=1}^{H_N,\tilde{P}} \Pi(|\gamma_{jh}-\gamma_{jh}^*| \leq \theta_N/\sqrt{\tilde{P}})$$
$$\geq \exp(-||\boldsymbol{\gamma}^*||_2^2 - \theta_N^2 H_N)(2\theta_N/\sqrt{\tilde{P}})^{H_N\tilde{P}} \gtrsim \exp(-M_N\theta_N^2 C_2),$$

for any $C_2 > 0$, where the first inequality follows from Assumption (E) and the last inequality follows from $H_N P \log(\sqrt{\tilde{P}}/2\theta_N) \prec M_N\theta_N^2$ (since $M_N\theta_N^2 \asymp M_N^{d/(d+2\xi)}$ while $H_N \prec M_N^{d/(d+2\xi)}$). ■

Appendix B. Proof of Lemma 3

Proof Denote $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N} = \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{B}$, $\widehat{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}^{\mathrm{T}}\boldsymbol{y}_{\boldsymbol{\Phi},N}$ and a sequence of random variables $\zeta_N = I(||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\widehat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2 \gtrsim \theta_N M_N^{1/2})$. Then,

$$\begin{aligned}\mathbb{E}^*(\zeta_N) &= P^*(||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\widehat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2 \gtrsim \theta_N M_N^{1/2}) \\ &= P^*(||\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^* + \boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},B,N}}\boldsymbol{\epsilon}||_2^2 \gtrsim \theta_N^2 M_N) \\ &\leq P^*(||\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2 + ||P_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\boldsymbol{\epsilon}||_2^2 \gtrsim \theta_N^2 M_N),\end{aligned}$$

where $\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}$ denotes the projection matrix corresponding to the matrix $\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}$. Note that

$$||\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2 \leq \boldsymbol{\eta}^{*\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}^{\mathrm{T}}\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^* \leq ||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2.$$

We then refer to equation (19) to see that $E_{\mathcal{U}}E_X||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2 \lesssim N\kappa_N M_N^{-2\xi/(2\xi+d)} \prec M_N\theta_N^2$. The above two facts together conclude that

$$\mathbb{E}_{\mathcal{U}}\mathbb{E}_X[||\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}\boldsymbol{\eta}^*||_2^2] \lesssim N\kappa_N M_N^{-2\xi/(2\xi+d)} \prec M_N\theta_N^2.$$

$\mathbb{E}^*(\zeta_N) \lesssim P^*(||\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\boldsymbol{\epsilon}||_2^2 \gtrsim \theta_N^2 M_N) = P^*(\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\boldsymbol{\epsilon} \gtrsim \theta_N^2 M_N)$.

Note that under P^*, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}})$, and, $e_{max}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}) \asymp 1$ (by Lemma 1). Also note that Lemma 1 of Laurent and Massart (2000) can be simplified to write $P^*(\chi_{p^*}^2 > x) \leq \exp(-x/4)$, for $x \geq 8p^*$. Further, $\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\boldsymbol{\epsilon}$ follows a χ^2 distribution with degree of freedom less than equal to $H_N\tilde{P} \prec M_N\theta_N^2 = M_N^{d/(d+2\xi)}$. Using all the above facts, we conclude that $E^*(\zeta_N) \lesssim \exp(-M_N\theta_N^2)$.

Next, for $\boldsymbol{\gamma} \in \mathcal{B}_N^c$, we show that $\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2^2 \gtrsim M_N\theta_N^2$. To see this, note that

$$\begin{aligned}\mathbb{E}_{\mathcal{U}}E_{\mathcal{X}}||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2^2 &= \mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}^{\mathrm{T}}\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right] \\ \asymp \kappa_N\mathbb{E}_{\mathcal{U}}\mathbb{E}_{\mathcal{X}}\left[(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{X}}_N^{\mathrm{T}}\tilde{\boldsymbol{X}}_N\boldsymbol{B}(\boldsymbol{\gamma}-\boldsymbol{\gamma}^*)\right] &\asymp N\kappa_N||\boldsymbol{\gamma}-\boldsymbol{\gamma}^*||_2^2/H_N \gtrsim M_N\theta_N^2,\end{aligned}$$

where the second line follows using similar calculations leading to equation (18).

Now, using the fact that $||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\widehat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}||_2 \geq -||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\widehat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2 + ||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2$, we obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\gamma}}(1-\zeta_N) &= P_{\boldsymbol{\gamma}}(||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\widehat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}^*||_2 \lesssim \theta_N M_N^{1/2}) \\ &= P_{\boldsymbol{\gamma}}(||\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\widehat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}\boldsymbol{\gamma}||_2 \gtrsim \theta_N M_N^{1/2}) \\ &\leq P_{\boldsymbol{\gamma}}(||\boldsymbol{P}_{\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},\boldsymbol{B},N}}\boldsymbol{\epsilon}||_2^2 \gtrsim \theta_N^2 M_N) \lesssim \exp(-M_N\theta_N^2),\end{aligned}$$

where the last inequality follows from simplifying the conclusion for Lemma 1 of Laurent and Massart (2000) (as is done before) and the fact that under $P_{\boldsymbol{\gamma}}$, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I})$. ■

Appendix C. Proof of Theorem 4

Proof Note that,

$$\begin{aligned}||\boldsymbol{w} - \boldsymbol{w}^*||_2 \leq ||\boldsymbol{w} - \tilde{\boldsymbol{w}}^* + \tilde{\boldsymbol{w}}^* - \boldsymbol{w}^*||_2 \leq ||\boldsymbol{w} - \tilde{\boldsymbol{w}}^*||_2 + ||\tilde{\boldsymbol{w}}^* - \boldsymbol{w}^*||_2 = ||\boldsymbol{w} - \tilde{\boldsymbol{w}}^*||_2 + ||\boldsymbol{\eta}^*||_2 \\ \lesssim ||\boldsymbol{w} - \tilde{\boldsymbol{w}}^*||_2 + P^{1/2}H_N^{-\xi} \asymp ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 H_N^{-1/2} + P^{1/2}H_N^{-\xi} \\ \asymp ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 H_N^{-1/2} + P^{1/2}M_N^{-\xi/(2\xi+d)},\end{aligned}$$

where $\tilde{\boldsymbol{w}}^*(\boldsymbol{u}) = (\sum_{h=1}^{H_N} B_{1h}(\boldsymbol{u})\gamma_{1h}^*, \ldots, \sum_{h=1}^{H_N} B_{\tilde{P}h}(\boldsymbol{u})\gamma_{\tilde{P}h}^*)^{\mathrm{T}}$, and the first inequality in the second line follows from the property of B-splines (Huang et al., 2004). The second expression in the second line follows from Lemma A.1 of Huang et al. (2004). Using the fact that $\tilde{P}^{1/2}M_N^{-\xi/(2\xi+d)} = O(\theta_N)$, we have $\left\{\boldsymbol{w} : ||\boldsymbol{w} - \boldsymbol{w}^*||_2 \geq \tilde{C}\theta_N\right\} \subset \left\{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 H_N^{-1/2} \geq C_{2w}\theta_N\right\}$, for some constant $C_{2w} > 0$.

Denote $\mathcal{B}_N = \left\{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 H_N^{-1/2} \leq C_{2w}\theta_N\right\}$. To prove the theorem, it is enough to establish

$$\mathbb{E}^*\Pi(||\boldsymbol{\gamma} - \boldsymbol{\gamma}^*||_2 H_N^{-1/2} \geq C_{2w}\theta_N | \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N}) \to 0, \text{ as } N \to \infty, \tag{20}$$

Note that,

$$\begin{aligned}&\mathbb{E}^*[\Pi(\mathcal{B}_N^c \,|\, \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N})] \leq \mathbb{E}^*\zeta_N + \mathbb{E}^*[\Pi(\mathcal{B}_N^c \,|\, \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N})(1-\zeta_N)1_{y_N \in \mathcal{A}_N^c}] + P^*(\mathcal{A}_N) \\ &= \mathbb{E}^*[\zeta_N] + \mathbb{E}^*\left[1_{\boldsymbol{y}_N \in \mathcal{A}_N^c} \frac{\left\{(1-\zeta_N)\int_{\mathcal{B}_N^c}\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right\}}{\left\{\int\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right\}}\right] + P^*(\mathcal{A}_N),\end{aligned} \tag{21}$$

where $\mathcal{A}_N$ is a set defined in the statement of Lemma 2 and ζ_N can be regarded as a sequence of random variables as defined in Lemma 3. By Lemma 2, $P^*(\mathcal{A}_N) \to 0$, as $N, M_N \to \infty$. Also, by Lemma 3, $\mathbb{E}^*\zeta_N \to 0$, as $N, M_N \to \infty$. To show (20), it remains to prove that

$$\frac{\mathbb{E}^*\left[1_{\boldsymbol{y}_N \in \mathcal{A}_N^c}\int_{\mathcal{B}_N^c}\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right]}{\left[\int\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right]} \to 0 \text{ as } N, M_N \to \infty.$$

To this end, we have

$$\mathbb{E}^*\left[1_{\boldsymbol{y}_N\in\mathcal{A}_N^c}\int_{\mathcal{B}_N^c}\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi_N(\boldsymbol{\gamma})d\boldsymbol{\gamma}\right] \leq \sup_{\boldsymbol{\gamma}\in\mathcal{B}_N^c}\mathbb{E}_{\boldsymbol{\gamma}}(1-\zeta_N)\Pi(\mathcal{A}_N^c)$$
$$\leq \exp(-C_{2w}M_N\theta_N^2),$$

where $\Pi(\mathcal{A}_N^c)$ is the prior probability of the set $\mathcal{A}_N^c$. The denominator

$$\int\{f(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma})/f^*(\boldsymbol{y}_{\boldsymbol{\Phi},N}|\boldsymbol{\gamma}^*)\}\pi(\boldsymbol{\gamma})d\boldsymbol{\gamma} \geq \exp(-C_1M_N\theta_N^2)$$

on $\mathcal{A}_N$, where C_1 is chosen so that $C_1 < C_{2w}$. Thus, $\mathbb{E}^*\Pi(\mathcal{B}_N^c \,|\, \boldsymbol{y}_{\boldsymbol{\Phi},N}, \tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N})1_{\boldsymbol{y}_N\in\mathcal{A}_N^c} \leq \exp(-(C_{2w}-C_1)M_N\theta_N^2) \to 0$, as $N, M_N \to \infty$. ■

Appendix D. Proof of Theorem 5

Proof For densities f_u and f^*, we have

$$h(f_u,f^*) = 1-\exp\left\{-\left(\sum_{j=1}^{\tilde{P}}\tilde{x}_j(\boldsymbol{u}_0)w_j(\boldsymbol{u}_0)-\sum_{j=1}^{\tilde{P}}\tilde{x}_j(\boldsymbol{u}_0)w_j^*(\boldsymbol{u}_0)\right)^2/8\right\}$$
$$\leq 1-\exp\left\{-\tilde{P}\sum_{j=1}^{\tilde{P}}\left(w_j(\boldsymbol{u}_0)-w_j^*(\boldsymbol{u}_0)\right)^2/8\right\}$$
$$\leq 1-\exp\left\{-\tilde{P}||\boldsymbol{w}(\boldsymbol{u}_0)-\boldsymbol{w}^*(\boldsymbol{u}_0)||_2^2/8\right\}$$

Then, $\mathbb{E}_{\mathcal{U}}[h(f_u,f^*)] \leq 1-\exp\left(-\tilde{P}||\boldsymbol{w}-\boldsymbol{w}^*||_2^2/8\right)$, by Jensen's inequality. Further,

$$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{U}}[h(f_u,f^*)|\tilde{\boldsymbol{X}}_{\boldsymbol{\Phi},N},\boldsymbol{y}_{\boldsymbol{\Phi},N}] = \left\{1-\exp\left(-\tilde{P}\tilde{C}^2\theta_N^2/8\right)\right\} + 2\Pi_N(||\boldsymbol{w}-\boldsymbol{w}^*||_2 \geq \tilde{C}\theta_N),$$

which implies

$$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{U}}[h(f_u,f^*)] \leq \left\{1-\exp\left(-\tilde{P}\tilde{C}^2\theta_N^2/8\right)\right\} + 2\mathbb{E}^*\Pi_N(||\boldsymbol{w}-\boldsymbol{w}^*||_2 \geq \tilde{C}\theta_N) \to 0$$

as $N, M_N \to \infty$, where the last expression followed by the conclusion of Theorem 4 and the fact that $\theta_N \to 0$ as $N, M_N \to \infty$. ■

References

D. Ahfock, W. J. Astle, and S. Richardson. Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*, 2017.

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

N. Ailon and B. Chazelle. The fast johnsonlindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39:302–322, 01 2009. doi: 10.1137/060673096.

R. Bai, M. R. Boland, and Y. Chen. Fast algorithms and theory for high-dimensional Bayesian varying coefficient models. *arXiv preprint arXiv:1907.06477*, 2019.

S. Banerjee. High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12:583–614, 2017.

S. Banerjee and G. A. Johnson. Coregionalized single- and multiresolution spatially varying growth curve modeling with application to weed growth. *Biometrics*, 62(3):864–876, 2006. doi: https://doi.org/10.1111/j.1541-0420.2006.00535.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00535.x`.

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2014.

R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In *International conference on machine learning*, pages 405–413. PMLR, 2014.

R. Bardenet, A. Doucet, and C. Holmes. On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

A. Bhattacharya, A. Chakraborty, and B. K. Mallick. Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042, 2016.

C. Biller and L. Fahrmeir. Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, 1(3):195–211, 2001.

N. Bliznyuk, D. Ruppert, C. Shoemaker, R. Regis, S. Wild, and P. Mugunthan. Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. *Journal of Computational and Graphical Statistics*, 17(2):270–294, 2008.

C. Brunsdon, A. S. Fotheringham, and M. E. Charlton. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4): 281–298, 1996. doi: https://doi.org/10.1111/j.1538-4632.1996.tb00936.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1996.tb00936.x`.

D. R. Burt, C. E. Rasmussen, and M. van der Wilk. Convergence of sparse variational inference in gaussian processes regression. *arXiv preprint arXiv:2008.00323*, 2020.

Z. Cai, J. Fan, and Q. Yao. Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956, 2000.

E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

R. Chen and R. S. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88(421):298–308, 1993.

S. Chen, Y. Liu, M. R. Lyu, I. King, and S. Zhang. Fast relative-error approximation algorithm for ridge regression. In *UAI*, pages 201–210, 2015.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

K. Clarkson and D. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63:1–45, 01 2017. doi: 10.1145/3019134.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008.

N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111:800–812, 2016. URL `http://dx.doi.org/10.1080/01621459.2015.1044091`.

C. De Sa, V. Chen, and W. Wong. Minibatch gibbs sampling on large graphical models. In *International Conference on Machine Learning*, pages 1173–1181, 2018.

S. K. Deshpande, R. Bai, C. Balocchi, J. E. Starling, and J. Weiss. Vcbart: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*, 2020.

E. Dobriban and S. Liu. A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*, 2018.

D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012.

A. O. Finley. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154, 2011. doi: https://doi.org/10.1111/j.2041-210X.2010.00060.x. URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2010.00060.x`.

A. O. Finley and S. Banerjee. Bayesian spatially varying coefficient models in the spbayes r package. *Environmental Modelling & Software*, 125:104608, 2020. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2019.104608. URL `https://www.sciencedirect.com/science/article/pii/S1364815219310412`.

A. O. Finley, S. Banerjee, and D. W. MacFarlane. A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, 106(493):31–48, 2011.

A. E. Gelfand, H.-J. Kim, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462): 387–396, 2003.

S. Ghosal, A. Van Der Vaart, et al. Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, 35(1):192–223, 2007.

G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

R. Guhaniyogi and S. Banerjee. Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. *Technometrics*, 60(4):430–444, 2018. doi: 10.1080/00401706.2018.1437474. URL `https://doi.org/10.1080/00401706.2018.1437474`. PMID: 31007296.

R. Guhaniyogi and S. Banerjee. Multivariate spatial meta kriging. *Statistics & probability letters*, 144:3–8, 2019.

R. Guhaniyogi and D. B. Dunson. Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514, 2015.

R. Guhaniyogi and D. B. Dunson. Compressed Gaussian process for manifold regression. *The Journal of Machine Learning Research*, 17(1):2472–2497, 2016.

R. Guhaniyogi and B. Sansó. Large multi-scale spatial kriging using tree shrinkage priors. *arXiv preprint arXiv:1803.11331*, 2018.

R. Guhaniyogi and A. Scheffler. Sketching in Bayesian high dimensional regression with big data using gaussian scale mixture priors. *arXiv preprint arXiv:2105.04795*, 2021.

R. Guhaniyogi, A. O. Finley, S. Banerjee, and R. K. Kobe. Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *Journal of agricultural, biological, and environmental statistics*, 18(3):274–298, 2013.

R. Guhaniyogi, C. Li, T. D. Savitsky, and S. Srivastava. A divide-and-conquer Bayesian approach to large-scale kriging. *arXiv preprint arXiv:1712.09767*, 2020a.

R. Guhaniyogi, C. Li, T. D. Savitsky, and S. Srivastava. Distributed Bayesian varying coefficient modeling using a gaussian process prior. *arXiv preprint arXiv:2006.00783*, 2020b.

R. Guhaniyogi, C. Li, T. Savitsky, and S. Srivastava. Distributed Bayesian inference in massive spatial data. *Statistical science*, 38(2):262–284, 2023.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

J. Z. Huang, C. O. Wu, and L. Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, pages 763–788, 2004.

Z. Huang. Near optimal frequent directions for sketching dense and sparse matrices. In *International Conference on Machine Learning*, pages 2048–2057. PMLR, 2018.

Z. Huang, J. Li, D. Nott, L. Feng, T.-P. Ng, and T.-Y. Wong. Bayesian estimation of varying-coefficient models with missing data, with application to the singapore longitudinal aging study. *Journal of Statistical Computation and Simulation*, 85(12):2364–2377, 2015.

S. Jeong and S. Ghosal. Unified bayesian asymptotic theory for sparse linear regression. *arXiv preprint arXiv:2008.10230*, 2020.

S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on signal processing*, 56(6):2346–2356, 2008.

M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, 2017. doi: 10.1080/01621459.2015.1123632. URL `http://dx.doi.org/10.1080/01621459.2015.1123632`.

M. Katzfuss and J. Guinness. A General Framework for Vecchia Approximations of Gaussian Processes. *Statistical Science*, 36(1):124 – 141, 2021. doi: 10.1214/19-STS755. URL `https://doi.org/10.1214/19-STS755`.

M. Kim and L. Wang. Generalized spatially varying coefficient models. *Journal of Computational and Graphical Statistics*, 30(1):1–10, 2021. doi: 10.1080/10618600.2020.1754225. URL `https://doi.org/10.1080/10618600.2020.1754225`.

A. Korattikara, Y. Chen, and M. Welling. Austerity in mcmc land: Cutting the metropolis-hastings budget. In *International conference on machine learning*, pages 181–189. PMLR, 2014.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

J. Lee, M. E. Kamenetsky, R. E. Gangnon, and J. Zhu. Clustered spatio-temporal varying coefficient regression model. *Statistics in medicine*, 40(2):465–480, 2021.

R. T. Lemos and B. Sansó. A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association*, 104(485):5–18, 2009.

D. Li and W. H. Wong. Mini-batch tempered mcmc. *arXiv preprint arXiv:1707.09705*, 2017.

J. Li, Z. Wang, R. Li, and R. Wu. Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The annals of applied statistics*, 9(2):640, 2015.

M. W. Mahoney. Randomized algorithms for matrices and data. *arXiv preprint arXiv:1104.5557*, 2011.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015. doi: 10.1080/10618600.2014.914946. URL `http://dx.doi.org/10.1080/10618600.2014.914946`.

M. Peruzzi, S. Banerjee, and A. O. Finley. Highly scalable Bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 117(538):969–982, 2022. doi: 10.1080/01621459.2020.1833889. URL `https://doi.org/10.1080/01621459.2020.1833889`.

T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, page 143152, USA, 2006. IEEE Computer Society. ISBN 0769527205. doi: 10.1109/FOCS.2006.37. URL `https://doi.org/10.1109/FOCS.2006.37`.

D. Seita, X. Pan, H. Chen, and J. Canny. An efficient minibatch acceptance test for metropolis-hastings. *arXiv preprint arXiv:1610.06848*, 2016.

W. Shen and S. Ghosal. Adaptive bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213, 2015.

E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18, 2005.

A. v. d. Vaart and H. v. Zanten. Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(Jun):2095–2119, 2011.

A. W. Van der Vaart, J. H. van Zanten, et al. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675, 2009.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B*, 50:297–312, 1988.

S. S. Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

B. Vidakovic. *Statistical modeling by wavelets*, volume 503. John Wiley & Sons, 2009.

H. Wang and Y. Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104(486):747–757, 2009.

L. Wang, H. Li, and J. Z. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103(484):1556–1569, 2008.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

D. C. Wheeler and C. A. Calder. An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2):145–166, 2007.

C. K. Wikle. Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, pages 107–118, 2010. Gelfand, A. E., Diggle, P., Fuentes, M. and Guttorp, P., editors, Chapman and Hall/CRC, pp. 107-118.

D. P. Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.

T.-Y. Wu, Y. Rachel Wang, and W. H. Wong. Mini-batch metropolis–hastings with reversible sgld proposal. *Journal of the American Statistical Association*, 117(537):386–394, 2022.

X. Yuan, P. Llull, D. J. Brady, and L. Carin. Tree-structure bayesian compressive sensing for video. *arXiv preprint arXiv:1410.3080*, 2014.

L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu. Recovering the optimal solution by dual random projection. In *Conference on Learning Theory*, pages 135–157, 2013.

L. Zhang, A. Datta, and S. Banerjee. Practical Bayesian modeling and inference for massive spatial data sets on modest computing environments. *Statistical Analysis and Data Mining: An ASA Data Science Journal*, 12(3):197–209, 2019. doi: https://doi.org/10.1002/sam.11413. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11413`.

L. Zhang, W. Tang, and S. Banerjee. Bayesian geostatistics using predictive stacking, 2024. URL `https://arxiv.org/abs/2304.12414`.

S. Zhou, L. Wasserman, and J. D. Lafferty. Compressed regression. In *Advances in Neural Information Processing Systems*, pages 1713–1720, 2008.